Weekly Report

Lu Junhua

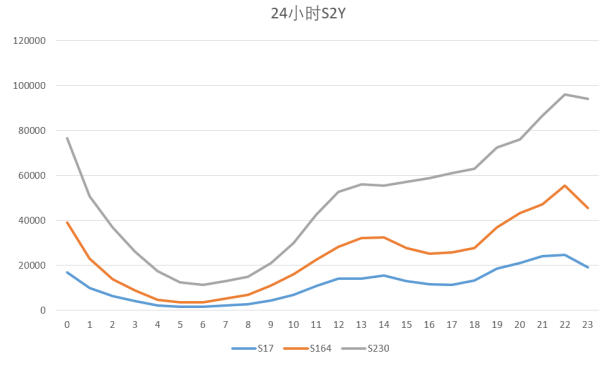
2015 年 11 月 30 日

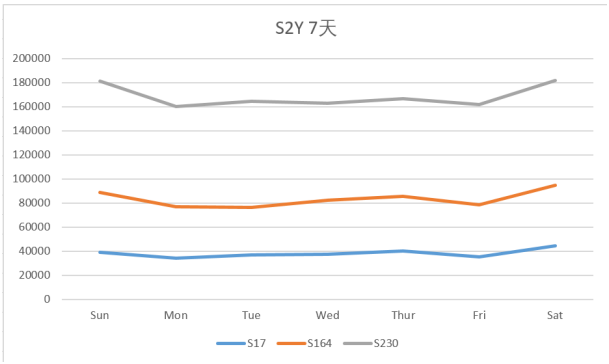
I wrote about 2000 words in the survey. Details can be seen below. We discussed and reconstruct the pipeline in the article. Ma gave much advices and I will hand on a more detailed version on Wednesday.

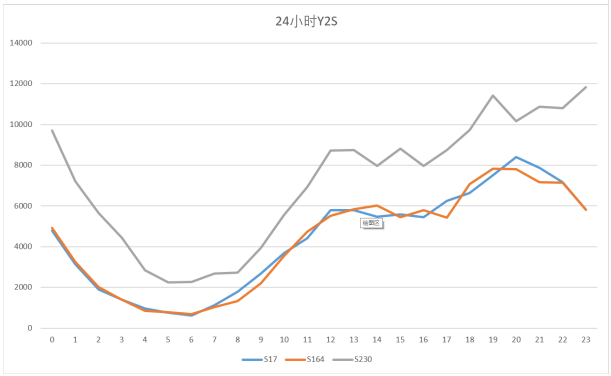
In netease games, still we are in exploration of data. We construct a bipartite network: If A and B bought the clothes of same clothesid, they are connected (on server 17, 164,230). At first, we hope to use built-in functions in Python networkx to compute the whole network in buy_shizhuang, but the ram is overflow. We had no no choice but to do preprocess and only choose server 17,164,230. The result is shown as below.

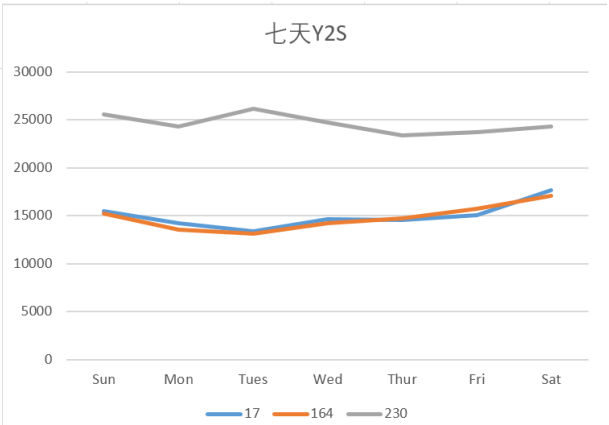
服务器id	记录总条数	赠送总次数	购买id数	连通分量数	密度	平均聚类系数
17	2824	237	1226	1	0.1685	0.8332
164	3963	352	1907	1	0.1555	0.8468
230	5376	500	3083	1	0.1838	0.8778
买过同一时装id的人之间有一条边, 这样构成的网络.						

And we also made statistics on silver - yuanbao exchange, aggregate by different intervals.(24 hours in a day, seven days in a week, count the times of exchange on server 17,164,230, and overall). I discovered that Y2S by hour curve resembles in shizhuang transaction by hour curve and server 230's Y2S is different from other 2 servers(because it's new?).









Next week, besides finishing the preliminary version of survey, I will do: read 30% of reviews on networks offered by Peng Taiquan; interview students; preparing presentation of my course; Do regression and multiple regression analysis on the present netease data.

Front.Comput.Sci.
DOI

REVIEW ARTICLE

A Survey on Predictive Visual Analytics

Junhua Lu (✉)[1]

1 Zhejiang University, Hangzhou 310058, China

Abstract A short abstract of up to 300 words written in one paragraph, clearly indicating the object and scope of the paper as well as the results achieved, should appear on the first page. It should be written using the abstract environment.

Keywords Up to 8 words separated by commas.

1 Introduction

placeholder This is reserved for Yuxin Ma
cite: [1–4]

2 A Conceptual Framework for Predictive Visual Analytics

What is predictive analytics? It is "the process of extracting information from large data sets in order to make make predictions and estimates about future outcomes". Most classification and estimation methods can be utilized to predict if properly applied, for example, regression, *k*NN, Support Vector Machines etc. However in the process of prediction, users may come across problems like inappropriate selection of features, incomprehensibility of models as black boxes. Therefore the involvement of human become important and visualization is a critical way to assist predictive analytics.

Overall:Efforts have been made to generalize the pipeline or part of the pipeline for predictive visual analytics. El-Assady et al. [5] summarize their work in VAST challenge 2013 as fig**??**. It is an contest of solving visual analytics and this year's tasks are predicting moving ratings and box offices. Visualization is adopted in different steps while predicting: filtering&weighting, automatic prediction and adjustment of models. This workflow enable users to interact with the system in feedback loops.

partial:anymoreBesides,some papers focused on specific stages of predictive model development. David Gotz et al. [6] propose a method to help improve model accuracy in iteratively using visualization to analyse outcome for model selection and then score the examples to identify problematic features and examples.

Gleicher M.[position paper] point out that comprehensibility is an import concern for predictive modeling. Based on his theory and inspired by [5], We propose a framework as in fig**??**. It consists of three main components: feature selection and generation, model generation & application and adjustment. Different visualization methods can depict characteristics of features and help users decide which to choose or combine. During model generation and selection, data mining algorithms and visualization techniques will be applied to achieve users' goals. After this stage, adjustment according to the previous outcomes is necessary for the next iteration of predictive visual analysis. Users' background, domain knowledge and experience can be included in all stages of predictive analysis using visualization techniques.

The rest of paper will be organized as follows: Sec 3 introduce different categories of common predictive visual analytics tasks; sec 4 introduces how visualization in helping feature selection& generation; sec 5 explains how to visualize model generation and application besides the automatic compute of computer; Adjustment of model and then return to the first step will be shown in sec 6. Challenges and future work will be listed in sec7.

Received month dd, yyyy; accepted month dd, yyyy

E-mail: ××××@×××.×××

3 Data Input

3.1 Divided by data types

Different categories of prediction data need different visualization and visual analytics techniques. Prediction tasks include temporal data prediction, spatial-temporal data, textual data, image data and user behaviour. We extracted the characteristics of each category as shown in Table1.

Table 1 A summary of data types in Predictive Visual Analytics

categories	Input	Output	ML/DM techniques
blank control	61.5	71.4	68.0
PLA	73.2	75.6	65.2
HA-PLA	54.4	78.6	62.4

There are many interesting and valuable application of all these tasks. For example the spread of epidemic[Integrating Predictive Visualization with the Epidemic Disease Simulation System (EpiSimS)], the prediction of trajectories, the diffusion of opinions and so. examples

3.2 Tasks

Also, we can divide by tasks.Before data can be used for further analysis, they should be processed under a sequence of operations: cleansing , transformation and integration. Many techniques have been developed in the field of machine learning, include detection of erroneous values, extraction of useful information, inference of type and schema match. [**?**].cite his related work as different steps of ...

Dirty data include a large range of kinds of data, they can be: missing data, uncertain data, wrong data, unusable data. [**?**] [**?**]. And for time-oriented data, there also exists problematic data like duplicated data, outdated data etc. [**?**]. Many factors could contribute to these problems like humans' incorrect manipulation, different source of data with different formats and conventions etc. Most of these need human intervention. Visualization ,in a way, can help identify and represent the quality of data, and users can interact with visualization techniques to correct the problems.

Kandel et al. [**?**] made a survey on the research directions in data wrangling, they propose a framework of visually data wrangling: Diagnosing data problems, living with dirty data(visualizing missing or uncertain data), transforming data and editing and auditing transformations. They also proposed Wrangler [**?**], an interactive system which combines users' operation and computers' automatic transformation of data.
More

4 Feature Selection and Generation

When data is high-dimensional, determining which features to choose is critical. Analyzers hopes remove features that are not informative and inoperative. There are many feature selection algorithms in statistics or machine learning. Statistical aggregation may hide the difference of features and Machine Learning algorithms often perform in a black-box manner, thus cause the low efficiency of feature selection.

Visualization techniques can help people step into the process of finding features. The process of feature selection and generation can also be visualized to iteratively implement.

5 Model Training

Model training follows the step of feature selection. Appropriate choose of features is important to the building of models ,and the result of model training in turn can loop back to the former process and refine the feature selection in an iterative way. Visualization techniques helps users involve into the process interactively.

Generally speaking, model training can be classified as black box manner and white box manner. Black box methods comprise input, output and parameters, run to a completion or return error, white white box methods could depict the computing process of model and help users better understand the relationship between input, output and parameters. It also proposed 4 strategies for spliting: best pure partitions, largest cluster partitioning, best complete partitioning and different distribution partitioning.

5.1 Blackbox

5.2 Whitebox

Decision tree and Support Vector Machine(SVM) are most frequently studied methods since they can be interactive with visualization or other techniques. Decision tree in particular, itself has an explicit tree-like structure which can be visualized, along with visualization techniques in other analysis steps, help users interactively build classifiers.

A key problem in constructing decision trees is the split of nodes. [**?**] supports bivariate split using 2D polygons on the

points in a 2d plane; [**?**] supports splitting into multiple intervals with interactions. BaobabView [**?**] makes steps further since it leverage different visualization techniques(like streamgraph, confusion matrix etc) to display the characteristics of data attribute, the correlations between attributes, results of different split values with the support of different interactions along with automatic algorithms. Users are able to directly modify the tree while viewing the structure of resulting tress of training data. [**?**] uses barcharts and piecharts to help build simple bayesian classifier, which is similar to decision trees building;

Caragea et al. [**?**] use projection-based tour method to help users navigate through high-dimensional spaces and better understand SVM. Ma et al. built [**?**] which helps users understand the primary structure of SVM and the classified results. Also they proposed a novel visualizational method of rule extraction.

6 Model Comparison and Selection

7 Model Validation

8 Application

9 Conclusion and Challenges

Visualization could make contributions in predictive analysis process at different stages, from data input to application. It helps users incorporate knowledge and experience into analysis, and under certain circumstance non-experts could also perform well with the aid of visualization.

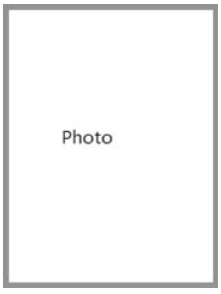
While predictive visual analytics have assist users analysis and predict in an interactive way, there still exists some problem remains to be solved. For example, most methods mentioned above can only cope with one single model, what if we had to choose between multiple models, how to arrange them in a way avoid collision? Under some circumstances, data input, feature selection and model generations may need very complicated computation and require large amount of resource and time, how to get instant feedback if in this critical situation. Also, designing an adapted predictive system is not an easy thing, it may need machine learning techniques itself to capture the characteristics of users with different backgrounds. Besides, most models are difficult for a user without much domain knowledge, i.e. most are in a black box way in the visual analytics as showed above. Whatąŕs more, the lack of connection between model and data also limit the model understanding and trust building in visualization. [12] And developing more techniques to assist users to understand the model will profoundly increase the accuracy of visually prediction.

Appendixes (if needed)

Appendix A

References

1. Verta O. Mastroianni C, Talia D. A super-peer model for resource discovery services in large-scale grids. Future Generation Computer Systems, 2005, 21(8): 1235–1248
2. Zhuge H. The Knowledge Grid. Singapore: World Scientific Publishing Co., 2004
3. Schlessinger D. Schaechter M. Bacterial toxins. In: Schaechter M, Medoff G, Eisenstein BI, eds. Mechanisms of microbial disease. 2nd ed. Baltimore: Williams and Wilkins, 1993, 162–175
4. Karger D. Ruhl M. Simple efficient load balancing algorithms for peer-to-peer systems. In: Proceedings of Sixteenth Annual ACM Symposium on Parallelism in Algorithm and Architectures. 2004, 36–40
5. El-Assady M. Jentner W. Stein M. Fischer F. Schreck T. Keim D. Predictive Visual Analytics - Approaches for Movie Ratings and Discussion of Open Research Challenges
6. Gotz D. Sun J. Visualizing Accuracy to Improve Predictive Model Performance



Please provide each author's biography here with no more than 120 words. The photo can be informal. Our journal prefers to exhibit an encouraging atmosphere. Please use a one that best suits our journal.